DOCUMENT RESUME

ED 131 100 TM 005 811

AUTHOR Calhoun, William Ford
TITLE Individualized Testing System. Report to Special Academic Programs. Improvement of Multiple Choice Testing.
PUB DATE 23 Oct 76
NOTE 53p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS *Computer Programs; Higher Education; *Multiple Choice Tests; Test Construction; *Testing Problems; Test Interpretation

ABSTRACT

This report documents (1) the problems inherent in multiple choice testing, (2) a solution to the problems, and (3) computer programs required by the solution. Problems of multiple choice testing include scheduling inflexibility, methodological inflexibility, cheating, inefficiencies of space and student interaction time, inefficiencies of instructor preparation time, uneven quality of examinations, and expense of preparation. The solution has four major features: (1) a special item file design, (2) a method of generating a large number of unique, equivalent measure tests, (3) published item files, and (4) a testing station independent of the classroom. A description and analysis of the practical implementation are also given. The implementation was highly regarded by the students (92 percent preference) as they believed that they learned more and performed better under this than under standard testing procedures. (HV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED131100

Individualized Testing System

Report to Special Academic Programs

"Improvement of Multiple Choice Testing"

16 October 1976

Redrafted 23 October 1976

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

William Ford Calhoun

Assistant Professor

Department of Biology

Virginia Polytechnic Institute and State University

Blacksburg, Virginia 24061

TM005 811

ABSTRACT

This report documents (1) the problems inherent in multiple choice testing, (2) a solution to the problems, and (3) computer programs required by the solution. It also describes and analyses a partial implementation of the solution. The implementation was highly regarded by the students (93% preference) as they believed that they learned more and performed better under it than under standard testing procedures.

ACKNOWLEDGEMENTS

Drs. Brian Davis, Ernie Stout, and A. Esen of the Biology Department provided valuable data and operational experience by using the individualized testing system in their classes. I am particularly indebted to Drs. Davis and Stout for valuable suggestions on improving the system and for suffering through, in good faith, many technical problems due to 'bugs' in the early versions of my computer programs and due to computer center conversions. Dr. Robert Frary of Learning Resources provided valuable suggestions, technical assistance, and encouragement. Without Dr. Frary's encouragement, I might have abandoned development of the system some time ago.

PROBLEM

Each quarter at Virginia Tech over 40,000 multiple-choice test sheets are processed by the Learning Resources Center. This number has been increasing at the rate of 10 to 20 percent per year since the service was established in 1972 and will probably exceed 50,000 sheets this quarter. For the most part, this testing is concentrated in a relatively small number of undergraduate courses with class sizes in excess of 100.

Under these circumstances use of machine processing followed by computer analysis and score reporting is virtually unavoidable. Given the resources available to faculty in charge of large sections, some with over 500 students, other methods of testing are simply too time consuming. Also, there is a further problem of maintaining uniformity of scoring when this work is distributed over a number of graduate assistants.

Quality and efficiency of multiple-choice testing is dependent upon a number of factors. However, the standard method of creating and administering multiple choice tests(A) has inherent deficiencies. A nonexhaustive list of deficiencies follows:

SCHEDULING INFLEXIBILITY

Students often miss tests, seek rescheduled tests, or perform poorly on tests for legitimate reasons. Unavoidable yet transitory personal problems and multiple tests scheduled on the same day in several courses are two of the more obvious reasons. In any case, either the conscientious instructor must assume extra administrative burden (validity of reasons often must be verified and alternative testing procedures must be maintained) or the hapless student must

---

(A) The standard method of multiple choice testing at VPI is: (1a) an instructor-typist interaction generates copies of a test or copies of scrambled versions of a test, or (1b) a computer program generates tests by selecting fixed items (each selection of an item produces the same information) from a file of fixed items, and (2) the tests are given to students at an assigned time and place(s).

suffer.

METHODOLOGICAL INFLEXIBILITY

Retesting required by mastery learning techniques and various other administrative and educational techniques can not be readily implemented with present testing systems.

CHEATING

While cheating by copying answers of an adjacent examinee has not received the press coverage of some other forms of cheating, the prevalence of this phenomenon has been documented in the University environment(B). Cheating on tests can be controlled by use of multiple forms of a test, careful control of seating arrangements, and rigid proctoring; but many instructors are unwilling to take such strenuous measures because they believe that student-instructor rapport will be adversely affected.

INEFFICIENCIES OF SPACE AND STUDENT-INSTRUCTOR INTERACTION TIME

Each class is allocated resources, principally in the form of X seat hours and Y instructor hours. The limits of student-instructor interaction are fixed by this allocation. Inclass testing consumes a significant portion of this allocation. For example, the typical lecture based course at VPI (3 contact hours per week for 9-10 weeks) requires 2-3 hours for inclass tests (excluding finals). That is, 7-11% of the allocated resources are consumed by tests. This use of time is not inconsequential as student-intructor interaction could be increased by that amount (presumably enhancing the student's education) or the student's stay at VPI could be reduced by that amount (saving the tax payers of Virginia considerable sums of money) if tests were removed from the classroom. For example, the equivalent of the VPI Introductory Biology series (3 contact hours for 3 quarters) could be taught as two, 4 hour courses.

(B) Frary, R. B and T. N. Tideman, Detection of Remarkable Similarity Between Multiple-choice Test Responses. Paper presented at annual meeting of American Psychological Association, Chicago, September, 1975.

INEFFICIENCIES OF INSTRUCTOR PREPARATION TIME AND UNEVEN QUALITY OF EXAMINATIONS

While tests are among the most powerful educational tools, preparation of each test is a creative effort which competes with research, administrative duties, lecture preparation, etc. for the instructor's time and energy. Since this competition occurs several times during the academic year, test preparation often comes out on the short end. Pushed for time, the instructor hurriedly creates a low quality test from scratch or uses essentially a previous test (which may be last quarter's last minute test). Education is poorly served in either case. Low quality tests frustrate the more conscientious students for these students are penalized relative to weaker students since chance plays a larger role in poor tests. Previous tests are a better measure of the students 'connections' than knowledge.

While some instructors make more of their efforts in this area than others, no instructor at VPI can readily analyse the quality and results of his tests in detail simply because the necessary use-statistics on test items are not readily available. Careful analysis of the results of tests is absolutely necessary in directing the instructor toward improvements in both his presentation and his testing procedure.

Uneven quality is also a problem with most computer based systems due to poor organization, lack of compactness of information, or lack of analytical features.

EXPENSE OF PREPARATION

The process of having a secretary type a test, the instructor proofread the test, and the secretary retype the test is time consuming and error prone. The investment increases tremendously if more than one form (scrambled version) of a test is required. While this expense is probably reduced by all computer based systems, the savings vary from system to system depending on design.

SOLUTION

A solution has been found (and in part demonstrated) that has 4 major features. The features are (1) a special item file design, (2) a method of generating a large number of unique, equivalent measure tests, (3) published item files, and (4) a testing station independent of the classroom. Each feature of the solution is discussed in turn.

ITEM FILE DESIGN

The item file was designed (1) to allow a large number of true and false responses for each item 'stem' (i.e., the question portion of the item), (2) to allow the inclusion of outline material, and (3) to allow the inclusion of response use-statistics. Use of these design features aid the instructor in creating, organizing, and evaluating his test items. They also aid the instructor in evaluating his presentations and the student's grasp of the study areas. Each design feature is discussed in turn.

Large Number of True and False Responses per Item Stem

Through the association of a large number of responses with each stem, the instructor reduces the need for redundant information, and the resultant concise information is much easier to access and evaluate. This multiple association also lends itself to efficient item creation as the major concept is usually embodied in the stem.

Outline Material

Outline material and comments are easily placed in the item file and are necessary for proper organization of an item file. A highly organized item file is absolutely necessary for even evaluation of both the student's grasp and the instructor's effectiveness. In fact, the item file should usually be more highly structured than other materials (e.g., lectures and text books) because it usually carries the same range of ideas as the presented material but at various levels of difficulty.

Response Use-Statistics

Although this feature is not available at present, it will soon be available and thus is discussed for completeness. Analysis will be accomplished by incorporating use-statistics of each response into the item file. These statistics will allow the instructor to determine which type of student missed the response and which type of student did not. These statistics are absolutely necessary if the instructor desires to measure his effectiveness, to measure the students grasp, and to measure the response's validity. Of course the 3 measures are confounded and no statistic will substitute for the instructor's experience and good judgement in deciding which measure is having the greater influence. But the fact remains that at present no instructor at VPI has access to the information needed to make that judgement.

In closing this section on the design of the item file, I hasten to restate that the item file is designed to allow organization and analysis. It encourages these characteristics by making them easily available, not by making them mandatory. In fact, even standard items (i.e., fixed items) can be entered into the item file with less effort than is customarily used in typing a test. This apparent laxity is purposeful, serving the larger role of allowing instructors to use the item file as if it were a standard item file. Thus, the instructor can grow into full use of the item file at his peculiar pace.

METHOD OF GENERATING UNIQUE, EQUIVALENT MEASURE EXAMS

As long as copies of only one test are available for use by a class during a testing period (e.g., a midterm) inflexibility must be the rule and cheating will be a correlary. That is, everyone involved must be at designated locations at a specific time, and in all probability someone will look at a neighbor's test. But the customary use of only one test results from administrative problems rather than conceptional ones. After all, with tests we are sampling the student's grasp of a study area through the use of a small sample of items. This small sample of items comes from the potentially infinite set of items stored in the instructor's brain. If an instructor could easily and randomly sample that set of items, he could generate a set of unique, yet equivalent measure tests. By unique, I mean that the tests would be sufficiently different that a student could not gain advantage by studying one of the tests, e.g., one taken by a friend, prior to taking another

of the tests for credit. By equivalent measure, I mean that any two tests from the set would provide equal measures of the student's grasp of the study area. With a set of unique, equivalent measure tests on a study area, the instructor could be quite flexible in administering tests and could be properly unconcerned with the common forms of cheating.

Thus, a part of the solution is a method of generating unique, equivalent measure tests. This is accomplished by randomly selecting a specified number of items from specified areas of an item file (areas corresponding to study areas) and randomly selecting true and false responses for each selected item. Equivalent measure of each test results from homogeneous areas of items and from the smoothing effect of random selection. Uniqueness results from the size of the item file. Given the compact design of the item file, the number of items needed for uniqueness of generated tests is well within the reach of an instructor.

PUBLISHED ITEM FILES

A sufficiently large (too large for a student to memorize), published item file would serve the immediate educational purpose of directing the student's study efforts and would reduce the administrative efforts required to maintain test security and to ensure equitable distribution of old tests (koofers at VPI) to all students.

TESTING STATION INDEPENDENT OF THE CLASS ROOM

A testing station independent of the classroom is necessary for efficiency of space and for flexibility. The testing station should consist of a comfortable room and a friendly human monitor. The monitor would check IDs, dispense appropriate tests to students as they enter the room, and dispense test keys to students who have completed their tests, providing the tested students with instant feedback. The station should be open at convenient times. Thus, a student would be free of time and other petty pressures that often adversely affect his performance. Operation of the station would provide economies of scale and skill. That is, many classes could use the same station, and the monitor's required skills would be minimal.

COMPUTER PROGRAMS

Certain computer programs are necessary for implementation of the solution. I have written three programs and a fourth remains to be written. The 3 existing programs are called ITSFMT, ITSANL, and ITSGEN. ITSUPD remains to be written. Each program is described in turn.

ITSFMT

ITSFMT reads through a raw (changed or recently created) file producing (1) an item file suitable for use by the other programs and (2) a statistics listing. This program allows the secretary to put less effort into creating or changing an item file than he would put into typing a test because ITSFMT does most of the work in formatting and because the statistics listing, with its descriptive statistics and error messages, makes detection and correction of logical errors easy. Figure 1 illustrates a small, raw file called FIG1. ITSFMT operated on FIG1 to produce item file FIG2 of figure 2 and simultaneously the statistics listing of figure 3. The remainder of this section is a detailed description of the raw file, statistics listing, and item file.

Raw Item File

The items entered in the raw file by the secretary have minimal format. The item ID is entered as one record, followed on subsequent records by the stem. Next the response ID is entered as one record, followed by the response on subsequent records. Other responses are entered in a like manner. The only other restrictions on entry of an item are that (1) records entered into the raw file must be less than 256 characters in length, (2) an item (stem + responses) and its comments must occupy less than 32,768 records in the item file where the maximum record length is 80 characters (longer records in the raw file are broken at word boundaries, the extra length being inserted as a new record), (3) an item must have a stem and at least 1 true and 1 false response, and (4) an item must be left justified at the 2nd column.

Column 1 is reserved for special characters. An item ID record must have a ':' in column 1 while a response ID record must have a '+' or '-' depending on the response's

truth value. An item ID may be any number of the form '#######.########' where '#' is a decimal digit (leading and trailing zeroes are not necessary). This range is provided to allow stability of old item IDs while allowing for insertion of new items and to allow latitude in choice of a systematic numbering system if one is desired. A response ID must be integers from 1 through 32,767. Both the item IDs and the response IDs (within an item) must be strictly monotonically increasing. Records with a '.' in column 1 and preceding a response ID record are considered comments. Similar records preceding an item ID record are considered outline material. The '.' records never appear on a test; they are strictly notes for the instructor.

Statistics Listing

If any logical errors exist in the raw file, ITSFMT will correct them or ignore the offending items. In either case, there will be no logical errors in the item file even though it may have fewer items than intended by the instructor. Thus, to aid the secretary in catching and correcting logical errors, both the type and location of logical errors in the raw file are documented in the statistics listing along with useful descriptive statistics. Examples of logical errors are improper numbering of items or responses, nonspecial characters in column 1, no responses for an item, etc. The statistics listing of figure 3 shows that raw file FIG1 has no logical errors. It also shows that the item beginning on record 3 of FIG1 has 3 more true responses and 5 more false responses indicated on the item ID record of FIG2 than indicated on the item ID record of FIG1. For the item beginning on the 24th record the numbers are 5 and 7. Both files have 53 records of 2 items.

Item File

An item file is an exact copy of a raw file (assuming no logical errors in the raw file) except (1) all records have been reduced to 80 characters or less and (2) additional fields may have been added and initilized. An item file has 3 fields on each item ID record that carry item descriptive numbers that are primarily for use by the other programs. These are the number of true and false responses, and the number of records occupied by the item and its comments. For example, in item file FIG2, item 139.557, the numbers are 3, 5, and 22, respectively. In addition to the item descriptive fields, several fields are set aside by ITSFMT to carry use-statistics once they become available. These fields are initilized to '0' if they are

undefined in the raw file. Each item ID record has a field reserved for the number of times the item appeared on a test, and each response ID record has 4 fields reserved for (1) the number of times a student reacted properly to the response (i.e., marked true responses and left false responses unmarked), (2) the number of times students reacted improperly to the response, (3) the sum of the standardized test scores of the properly reacting students, and (4) the sum of the standardized test scores of the improperly reacting students.

After items have been used on tests, the said use-statistics would be accumulated from the students answer sheets and added to the item file. ITSUPD would handle this chore and will be discussed later. However, it serves my presentation to show how item file FIG2 might appear with use-statistics. Figure 4 is item file FIG2 with simulated use-statistics. This file is now called FIG4.

In closing this section on ITSFMT, I restate that ITSFMT allows a secretary to create an item file with less effort and in less time than is required to type a test. This detailed discussion is presented for documentation of the program and is not necessary knowledge for its use. In fact, the structure of the entered item (the only one the secretary has to understand) is sufficiently intuitive that few errors will be made and the most likely ones will be corrected by the program. ITSFMT is easily operated. It needs to know where the raw file is located and where the item file is to be placed. The statistics listing is written on SYSPRINT.

ITSANL

Even though the item file is quite attractive to the computer programs and ensures execution efficiencies, an exact listing of the item file is definately not for human use. The tremendous detail of an exact listing prevents an overview of the item file, and the cumulative statistics prove awkward in analysis. Thus, ITSANL was written to provide two listings for the instructor: (1) a contents listing that contains outline material and item IDs and (2) an analytical listing that contains easily understood statistics instead of the cumulative statistics. Figures 5 and 7 were simultaneously produced by ITSANL from item file FIG4. Figure 5 is the contents listing and, figure 7 is the analytical listing. However, since item file FIG4 is so small that an overview is obvious, to carry the point I provide a contents listing from a larger item file (figure

6).

From the analytical listing the instructor can learn for each response (1) the number of times it appeared on a test, (2) the percentage of students who reacted properly to it (i.e., marked a true response or left a false response unmarked), (3) the average standardized score of the proper reactors, and (4) the average standardized score of the improper reactors. For example, response 1 of item 139.557 (figure 7) was on 424 tests; 20% of the students reacted properly to it; the average standardized score of the 20% was 79; and that of the other 80% was 45. Through careful reflection on this analytical information the instructor can measure his performance in each area of presentation, the students' grasp of each area, and each response's validity.

One should contrast the ease of understanding the information carried in figure 7 with that of figure 4. In most cases, the instructor would have no reason to see exact listings of his item file, and he would work with contents and analytical listings only.

ITSANL is easily operated. It needs to know where the item file is located and where the contents and analytical listing are to be placed.

ITSGEN

ITSGEN generates (1) a specifications listing and (2) a set of tests. The specifications listing lists the specifications used in generating the tests and any errors encountered.

The instructor must provide ITSGEN with one group of specifications for each set of tests to be generated. Three types of specifications are required for generation of a set of tests; other specifications are available to allow greater control over the generation of a set of tests and to allow for the production of auxiliary information. Specifications are of the form 'keyword = value;', and the last specification of a group must be followed by 'END;'. Each specification must be less than 32,768 characters in length. Table 1 is a list of all keywords, their use, and default values. The tests of figure 10 and the specifications listing of figure 9 resulted from the operation of ITSGEN on the group of specifications in figure 8. A discussion of this example will serve to illustrate the use of specifications.

Two selections ('SELECTIONS = 2;') were specified. A selection is a test generated by randomly picking items from the pool of specified items and randomly picking responses from the chosen items. In this example, two items were randomly picked from the items with IDs between 100 and 200, inclusively ('REQUEST = 2 100 - 200;'), and, by default, up to 1 true and 4 false responses were randomly picked from each chosen item. Thus, each selection was a unique test. Two versions ('VERSIONS = 2;') of each selection were specified. A version is a test generated by randomizing the order of each item's chosen responses. Thus, the two versions of each selection had the same items and responses but the responses were scrambled. Most instructors would not ask for both versions and selections. Many selections and 1 version (multiple tests) would be the rule for individualized testing, and many versions and 1 selection (multiple forms) would be the rule for inclass testing. I specified both to show the generality of ITSGEN. The tests were 80 characters wide ('LINESIZE = 80;') and had a title on the first page of each test ('TITLE = SPECIFICATION EXAMPLE;'). Item file FIG4 contained the requested items ('ITEM_FILE = FIG4;'), and the generated tests were written on file FIG10 ('TEST_FILE = FIG10;').

Figure 11 is another example of specifications used to generate a test. Figure 12 is the specifications listing, figure 13 the test, and figure 14 the test key. In this example, 4 item files were searched for 18 peculiar items. This is a realistic example as 240 copies of the generated test and key were printed and handed out to students as a study guide. This effort cost about 10 computer dollars and 10-20 human minutes even though no special effort was made toward efficiency.

A specification listing has up to 4 sections. One is essentially a relisting of the specifications. Any logical error in the specifications would be caught here, an appropriate error message printed, and execution on that group of specifications would cease. The second section lists all specifications obtained. The third section lists the items requested ('POSSIBLE ITEMS') in a sorted, parsed form. Any logical errors in an item request would be caught at this point, an appropriate error message printed, and execution on that group of specifications would cease. The fourth section lists the number of requested and available items. Should insufficient items be available, an error message would be printed and execution on that group of specifications would cease.

ITSGEN is easily operated. It needs to know where the

item files and specifications are located and where the tests and requested auxiliary information are to be placed.

ITSUPD

This program is not completed. It will combine the information in the analysis file (an auxiliary information file that carries each item and response ID of each generated test) with the students' test results and compute use-statistics. The results will be used to update the use-statistics of the item files.

IMPLEMENTATION OF SOLUTION

While the solution seemed logical, I realized that 'logical' solutions do not necessarily function as expected. Thus, to demonstrate its feasibility and to gain operational information, the solution was implemented on a small scale. This implementation is labeled ITS (individualized testing system).

PROCEDURE

Four instructors in 4 different classes were involved in this implementation of the solution (Fall 1975 and Winter 1976 Introductory Physiology and Winter 1976 and Spring 1976 Introductory Genetics with 173, 95, 163, and 97 students, respectively). Two major compromises were necessary.

(1) The solution requires a testing station available at convenient hours, e.g., those of a library. This would allow the students wide latitude in scheduling their tests and would allow the instructor to test and retest in any pattern or frequency desired. However, the scale and resources were too small to justify operating the testing station as envisioned. Thus, the instructors agreed to use only one pattern of testing. This was essentially a test-retest pattern at standard testing periods. A student could take 1 test during the first week and 1 test during the second week of a testing period. If the student took both tests, only the second test (retest) counted (finals were exceptions in 2 of the classes as the maximum of the test-retest counted). Grades were posted after the first week so that the students could make informed decisions on taking a second test (2 classes posted letter grades, 1 class posted scores, and 1 class posted adjusted scores). The testing station was generally open 12 hours a week (3 hours in the evening on each of 4 consecutive days).

(2) The solution requires item files sufficiently large for publication (too large for memorization). These publications would aid a student in directing his study efforts and would antiquate most efforts undertaken in the name of test security and equitable distribution of old tests. However, sufficient item files were not available. Thus, as a study guide, copies of a test and its key were distributed a week before the testing period. The security of the item files was maintained by retaining all tests at

the testing station.

A prototype of ITSGEN was used to generate sets of unique tests for each testing period. Two instructors created item files from scratch while the other 2 modified existing item files. Three of the classes used the system for all testing periods: midterms, endterms, and finals. One of the classes used it only for its endterm, using essay tests for the midterm and final. All students were given instructions on the use of the testing station (figure 15) a week before the first testing period. The protocol at the testing station was: the incoming student's ID was checked against a class roll and he was given the appropriate test, the tested student exchanged his answer sheet for the key to his test and studied his test and key, and the departing student returned his test and key to the monitor.

ANALYSIS

While the solution is very broad, ITS was necessarily tentative and narrow. Thus, an analysis of all aspects of the solution is not possible from ITS (most benefits of the solution are self-evident anyway). However, from student's surveys, grades, and conversations, an analysis of the educational effects of some aspects of the solution can be made; and from experience with this implementation, I can comment on cost effectiveness.

Students in 3 of the classes were surveyed after their first ITS testing period. Figure 16 is the survey and results. The numbers preceding the responses are the percentage (rounded to integers) of the students who considered the item and marked that response. The three columns represent the 3 classes surveyed, and the classes are referenced 'left', 'middle', and 'right' depending on the position of their column. A '_' means that the response was not on the survey given to that class. The surveys probably had the following validity relationship: 'right' > 'middle' > 'left'. The primary reasons for this ranking were (1) the number and percentage of students responding to the surveys (62% of 97, 'left'; 65% of 95, 'middle'; and 76% of 163, 'right') and (2) the propriety of the test items. The instructor of the 'right' class used his own item files, and they closely paralled his lectures; the instructors of the 'middle' class and 'left' class primarily used item files created by others, and they encountered and coped with divergence of lectures and item files to different degrees -- the instructor of the 'middle' class was more active and successful.

Impropriety of the test items given to the 'left' class is further supported by the fact that the 'left' class performed much poorer at a comparable testing period than the 'right' even though essentially the same item files were used to generate the tests for both classes (mean = 60.6, s.e. = 1.62 for 'right' and mean = 50.8, s.e. = 1.84 for 'left'). In addition, from conversations with students, I believe that students of the 'left' class were led to blame the 'computer' for the impropriety of their test items. Thus, I restrict my discussion of the survey to the 'right' class, though there was general agreement of all three surveys.

Survey comments (item 23) were divided into 3 classes: positive (figure 17), negative (figure 18), and irrelevant. Comments considered irrelevant were "I like the course but the book is terrible.", "Large computer sheet was very nice to have.", and statements on imaginary, corrected, or uncorrectable technical problems.

Educational Aspects, Positive

The reaction to ITS was so overwhelmingly positive that a sophisticated analysis was not required. Students were pleased with ITS because they believed that they learned more and performed better under it than under standard testing procedures. This opinion was independent of the pattern of testing (test-retest) since 80% of the students (item 19) would have perferred ITS to inclass testing even if retesting were not available. The benefits offered by the testing station were probably responsible for this acceptance. That is, 67% (item 13) thought choosing their test day and 81% (item 17) thought the atmosphere of the testing station contributed positively toward their performance. Even though 80% is a high acceptance rate, it is probably a low estimate of a true acceptance because (1) the testing station was constantly moved due to room scheduling problems, (2) some of the rooms were poorly designed for this purpose, (3) the monitor proved to be less than adequate on more than one occasion, and (4) the 4 day testing interval probably provided less scheduling flexibility than needed because many students have multiple tests during the week of a standard testing period. Although the testing intervals were functionally short, the students made the most of them and distributed themselves over the days (item 7) for various reasons: other tests (79%, item 8), study time (59%, item 9), personal reasons (35%, item 11), and illness (3%, item 10).

While the solution does not require a test-retest pattern of testing, this is an example of a pattern that could not be readily used with standard testing procedures. Thus, the students' acceptance of and performance under this pattern were measures of the solution. Acceptance of the retest option was high as the average student took 1.8 of the 3 possible retests (20%, 33%, and 32% of the 'right' class took retests at 1, 2, and 3 testing periods, respectively). 88% of the retesters (item 16) believed they learned more as a result of retesting, and the average retester significantly improved his score (for the 'right' class the average score changes for the 3 testing periods were 14%, s.e. = 1.3; 11%, s.e. = 1.3; and 7%, s.e. = 0.99).

Retesters improved their scores whether they studied as usual for the first test or studied in anticipation of the retest. An estimated 49% (based on surveyed students who gave their IDs and marked item 15) of the retesters claimed that they studied as usual for the first test, and they increased their score by 13% (s.e. = 3.2). One member of this group calculated that her class grade under a standard testing procedure would have been her usual C, but by correcting weaknesses observed in the first tests and taking all retests, her final grade was an A. An estimated 51% of the retesters claimed that they studied in anticipation of the retest, and this group increased its score by an estimated 17% (s.e. = 2.8).

Based on conversations with students, I believe that those anticipating the retest can be divided into at least 3 groups: (1) those who found it 'inconvenient' to prepare for the first test, (2) those who had genuine scheduling problems not solved by the 4 day testing interval, and (3) those who found that practicing on the first and concentrating on the second test was an effective method of study. Theoretically, a student maximizes his grades by preparing for the first test and viewing the second as a 'backup', but many of the 'practice-concentrate' students, who were 'good' students, believed that their method was best for them. Thus, the student's reactions to the test-retest pattern were highly varied and personalized, and the various personalities seemed to use it to benefit their educations.

In closing this section, I restate that the reaction to ITS was overwhelmingly positive. This fact is summarized in the following: (1) the grade shifts of retesters (table 2), (2) the 93% preference for ITS (item 18, figure 16), and (3) the student's positive comments (figure 17).

Educational Aspects, Negative

Possible negative aspects of ITS were surveyed with items 21 and 22 of figure 16. 16% (item 21) of the students felt that there were substantial variations in difficulty of the tests. Having seen these tests, I do not believe this to be true. In any case, the following experiment was done to check the systems ability to generate equivalent measure tests. The same item files used to generate the suspect tests were used to generate 2 versions of 2 selections with 25 items each. The 2 selections had only 3 common items with the same true responses. Copies of the tests were used for inclass testing (Brian Davis' Genetics, Fall 1976). The average score for the 212 students was 17.66 (s.e. = 0.28), the average scores for the 2 versions of selection 1 were 18.45 (s.e. = 0.58) and 17.21 (s.e. = 0.51) and those for the 2 versions of selection 2 were 17.39 (s.e. = 0.57) and 17.54 (s.e. = 0.58). This experiment does not indicate a difference in difficulty between the selections. Thus, the 16% was probably an estimate of frustration or rumor (at most a student saw 2 tests) rather than a measure of variation in test difficulty.

51% (15% + 36%, item 22) of the students believed that the 2 week testing period interfered with learning new material to some degree. This response is difficult to interpret because ITS does not require such an interference. Perhaps the 51% is a measure of the student's problems in scheduling their study time.

The negative comments are in figure 18. While I sympathize with the students' desire to have the instructor present at the testing station and while the solution does not require that the instructor or his aids be absent, I believe that testing is better served if the student is 'on his own'. This decreases noise and confusion in the testing room and eliminates the advantages that might otherwise be gained by the more persistent students.

Cost Effectiveness

Though little quantitative data exist, the following observations can be made: (1) the secretaries spent less time entering items than they would have spent typing the tests, (2) the rooms used as testing stations would have otherwise been unused, (3) the instructors 'gained' a class period at each testing period, (4) monitors were relatively cheap (we used a teaching assistant, but a work study student could be used because trustworthiness is the only

needed skill), and (5) tests cost from 4 to 20 computer cents per copy. A careful accounting would probably show that this implementation was cheaper to administer than standard testing procedures even though this implementation had start-up costs, could not take advantage of scale, and employed computer programs (prototypes of those discussed in this report) that were relatively inefficient and awkward to use.

CONCLUSION

This report has dealt with the solution at the level of the individual class. It shows that all problems inherent in standard multiple choice testing are addressed by the solution, and it shows that the solution is popular with students for educationally sound reasons. However, the benefits of testing stations with sets of unique, equivalent measure tests (not necessarily all multiple choice) could have many 'higher' applications. Following are 3 possibilities.

(1) Students could readily obtain credit (even grades) for a course by taking the final for that course. The departments would encounter no additional expense, yet students using this method would be subject to the same standards as students who took the course. Precocious entering freshmen, some transfer students (those with certain credit transfer problems), and some bright upper classmen might eliminate several lower level course requirements, allowing them to complete college sooner or to take more free electives. This method might be especially rewarding if current item files and study guides were maintained at high school libraries, and testing stations were available at freshman orientation. Perhaps students from 'better' high schools could eliminate much of the freshman year requirements.

I realize that students can presently exempt certain classes by examination. However, the present procedure does not ensure that the student has the specific facts needed to complete subsequent courses of a required series.

(2) With little or no effort, departments could require graduating seniors to take a spectrum of their tests. The test results, coupled with those of appropriate national tests, would be a measure of the students' total education, and thus should help the departments find weaknesses in their programs.

(3) Classes with multiple sections could be readily and evenly tested, and the relative strengths and weaknesses of the instructors could be readily ascertained. These instructor diagnostics would be much more meaningful than those presently obtained by VPI's student survey method as these would be based on actual student performance.

In addition to the 'higher' applications resulting from a testing station with sets of unique, equivalent measure tests, the item file design with its emphasis on compactness, organization, and response analysis should aid in the production of high quality testing material. The emphasis on publication should encourage distribution of this material. Indeed, authors of this material could find themselves in a market situation similar to that of text books. Thus, the more talented producers of item files would be rewarded for their efforts, and the entire education community would share the benefits at a nominal cost.

In closing this report, I stress that I have listed only a <u>few</u> 'higher' applications of the solution. No doubt the solution has many other far reaching applications as the production of large sets of high quality, unique, equivalent measure tests reduces most time-space-effort constraints on testing.

```
.TWO SAMPLE ITEMS FOR A DEMONSTRATION OF THE COMPUTER PROGRAMS
.Item from a genetics item file of 550 items.
:139.557
 The following, listed in order of dominance, are 4 alleles in rabbits:  c+,
 colored; c(ch), chinchilla; c(h) himalayan;
 c, albino.  A colored rabbit crossed to a colored rabbit produced 3 colored: 1
 chinchilla.
 The parental genotypes could have been
+1
 c+ c(ch) X c+ c(ch)
+2
 c+ c(ch) X c+ c(h)
-3
 c+ c(ch) X c(ch) c(ch)
-4
 c+ c(ch) X c(ch) c(h)
-5
 c+ c(h) X c(ch) c(ch)
+6
 c+ c(ch) X c+ c
-7
 c+ c X c(ch) c(ch)
-8
 none of the other answers are correct
.Item from a physiology item file of 242 items.
:159.5
 An amino acid residue that contributes to H bonding between amino acid side
 chains (R groups)
+1
 serine
+2
 tyrosine
.comments can be placed before any response
+3
 threonine
+4
 aspartic acid
+5
 glutamic acid
-6
 proline
-7
 tryptophan
-8
 isoleucine
-9
 leucine
-10
 valine
-11
 alanine
-12
 glycine
```


Figure 1. A raw file with 2 items, 139.557 and 159.5. Note that column 1 is reserved for special characters.

```
.TWO SAMPLE ITEMS FOR A DEMONSTRATION OF THE COMPUTER PROGRAMS   761004234915000
.Item from a genetics item file of 550 items.
:    139.55700000          0    3    5   22
 The following, listed in order of dominance, are 4 alleles in rabbits:  c+,
 colored; c(ch), chinchilla; c(h) himalayan;
 c, albino.  A colored rabbit crossed to a colored rabbit produced 3 colored: 1
 chinchilla.
 The parental genotypes could have been
+    1          0          0          0          0
 c+ c(ch)  X c+ c(ch)
+    2          0          0          0          0
 c+ c(ch)  X c+ c(h)
-    3          0          0          0          0
 c+ c(ch)  X c(ch) c(ch)
-    4          0          0          0          0
 c+ c(ch)  X c(ch) c(h)
-    5          0          0          0          0
 c+ c(h)  X c(ch) c(ch)
+    6          0          0          0          0
 c+ c(ch)  X c+ c
-    7          0          0          0          0
 c+ c X c(ch) c(ch)
-    8          0          0          0          0
 none of the other answers are correct
.Item from a physiology item file of 242 items.
:    159.50000000          0    5    7   28
 An amino acid residue that contributes to H bonding between amino acid side
 chains (R groups)
+    1          0          0          0          0
 serine
+    2          0          0          0          0
 tyrosine
.comments can be placed before any response
+    3          0          0          0          0
 threonine
+    4          0          0          0          0
 aspartic acid
+    5          0          0          0          0
 glutamic acid
-    6          0          0          0          0
 proline
-    7          0          0          0          0
 tryptophan
-    8          0          0          0          0
 isoleucine
-    9          0          0          0          0
 leucine
-   10          0          0          0          0
 valine
-   11          0          0          0          0
 alanine
-   12          0          0          0          0
 glycine
--------------------
```


Figure 2. Item file FIG2 produced from raw file FIG1 by ITSFMT. FIG2 is in proper form for use by the other computer programs.

```
IO SAMPLE ITEMS FOR A DEMONSTRATION OF THE COMPUTER PROGRAMS   761004234915000

RUE RESPONSES CHANGED BY 3 & FALSE CHANGED BY 5 FOR FOLLOWING ITEM
    3  ':139.557'
RUE RESPONSES CHANGED BY 5 & FALSE CHANGED BY 7 FOR FOLLOWING ITEM
   24  ':159.5'

TATISTICS:
INPUT    ITEMS =     2   RECORDS =    53
OUTPUT   ITEMS =     2   RECORDS =    53
```

---

igure 3. ITSFMT produced this statistics listing and item file FIG2 multaneously from raw file FIG1. Since FIG1 had no logical errors, ly statistics appear in the listing.

```
.TWO SAMPLE ITEMS FOR A DEMONSTRATION OF THE COMPUTER PROGRAMS   761004234915000
.Item from a genetics item file of 550 items.
:    139.55700000      1273    3    5   22
 The following, listed in order of dominance, are 4 alleles in rabbits:  c+,
 colored; c(ch), chinchilla; c(h) himalayan;
 c, albino.  A colored rabbit crossed to a colored rabbit produced 3 colored: 1
 chinchilla.
 The parental genotypes could have been
+    1        85       339      6715     15255
 c+ c(ch) X c+ c(ch)
+    2        84       316      6804     14536
 c+ c(ch) X c+ c(h)
-    3       190        48     11400      2160
 c+ c(ch) X c(ch) c(ch)
-    4       219        51     13578      2193
 c+ c(ch) X c(ch) c(h)
-    5       201        54     11859      2376
 c+ c(h) X c(ch) c(ch)
+    6        89       360      7298     15480
 c+ c(ch) X c+ c
-    7       205        45     12195      1935
 c+ c X c(ch) c(ch)
-    8       204        56      8832      3584
 none of the other answers are correct
.Item from a physiology item file of 242 items.
:    159.50000000      1500    5    7   28
 An amino acid residue that contributes to H bonding between amino acid side
 chains (R groups)
+    1       120       180      7440      9000
 serine
+    2       130       195      8190      9555
 tyrosine
.comments can be placed before any response
+    3       110       165      6710      8415
 threonine
+    4        58       232      4350     11600
 aspartic acid
+    5        62       248      4464     12400
 glutamic acid
-    6       192        22      9792      1078
 proline
-    7       154        66     11088      1980
 tryptophan
-    8       180        20      9360       980
 isoleucine
-    9       212        23     10812      1104
 leucine
-   10       165        29      8250      1450
 valine
-   11       174        38      8700      1900
 alanine
-   12       182        43      8736      2193
 glycine
--------------------
```

Figure 4. Item file FIG2 with simulated use-statistics.

TWO SAMPLE ITEMS FOR A DEMONSTRATION OF THE COMPUTER PROGRAMS 761004234915000

TWO SAMPLE ITEMS FOR A DEMONSTRATION OF THE COMPUTER PROGRAMS 761004234915000

Item from a genetics item file of 550 items.
: 139.55700000 1273 3 5

Item from a physiology item file of 242 items.
: 159.50000000 1500 5 7

Figure 5. Contents listing for item file FIG4.

FCPHA, FORD CALHOUN, INTRO PHYSIOLOGY, BIOL 3111 760921140105000

FCPHA, FORD CALHOUN, INTRO PHYSIOLOGY, BIOL 3111 760921140105000

AQUEOUS SOLUTIONS

1. structure of water				
:	1.00000000	0	2	12
:	2.00000000	0	3	5
:	3.00000000	0	1	4
:	4.00000000	0	5	5
2. chemical equilibria				
:	5.00000000	0	2	7
:	6.00000000	0	3	4
:	7.00000000	0	1	2
3. pH calculations				
3.1. strong acids added				
:	8.00000000	0	1	4
:	9.00000000	0	1	4
:	10.00000000	0	1	4
:	11.00000000	0	1	4
3.2. strong base added				
:	12.00000000	0	1	4
:	13.00000000	0	1	4
:	14.00000000	0	1	4
:	15.00000000	0	1	4
3.3. compute |H+|				
:	16.00000000	0	1	4
:	17.00000000	0	1	4
:	18.00000000	0	1	4
:	19.00000000	0	1	4
3.4. for weak acids				
:	20.00000000	0	1	4
:	21.00000000	0	1	4
:	22.00000000	0	1	4
:	23.00000000	0	1	4
:	24.00000000	0	1	4
:	25.00000000	0	1	4
:	26.00000000	0	1	4

FCPHA, FORD CALHOUN, INTRO PHYSIOLOGY, BIOL 3111 760921140105000

3.5. pKb from ionization constant				
:	27.00000000	0	1	4
:	28.00000000	0	1	4
:	29.00000000	0	1	4
4. buffers				
:	30.00000000	0	2	4
:	31.00000000	0	2	4
:	32.00000000	0	1	4
:	33.00000000	0	1	4
:	34.00000000	0	1	4
:	35.00000000	0	1	4
:	36.00000000	0	5	11
THERMODYNAMICS				
:	38.00000000	0	1	4
:	39.00000000	0	4	9
:	40.00000000	0	7	7
:	41.00000000	0	3	4
:	42.00000000	0	5	12
:	43.00000000	0	2	4
:	44.00000000	0	3	4
:	45.00000000	0	3	4
:	46.00000000	0	3	11
CHEMICAL COMPONENTS OF CELL				
1. lipids				
:	47.00000000	0	5	13
:	48.00000000	0	6	6
:	49.00000000	0	6	6
2. carbohydrates				
:	50.00000000	0	2	4
:	51.00000000	0	4	6
:	52.00000000	0	3	9
:	53.00000000	0	4	5
:	54.00000000	0	2	5
:	55.00000000	0	3	4
:	56.00000000	0	4	4
:	57.00000000	0	5	6
:	58.00000000	0	5	5

PCPHA, FORD CALHOUN, INTRO PHYSIOLOGY, BIOL 3111 760921140105000

3. amino acids

:	59.00000000	0	1	2
:	60.00000000	0	1	2
:	61.00000000	0	1	2
:	62.00000000	0	1	4
:	63.00000000	0	1	4
:	64.00000000	0	1	4
:	65.00000000	0	1	4
:	66.00000000	0	1	5
:	67.00000000	0	1	4
:	68.00000000	0	2	5
:	69.00000000	0	2	5
:	70.00000000	0	2	4
:	71.00000000	0	3	5
:	72.00000000	0	1	5
:	73.00000000	0	1	5

4. proteins

4.1. structure

:	74.00000000	0	4	7
:	75.00000000	0	5	6
:	76.00000000	0	7	6
:	77.00000000	0	6	6
:	78.00000000	0	1	19
:	79.00000000	0	3	7
:	80.00000000	0	4	8
:	81.00000000	0	1	19
:	82.00000000	0	5	7
:	83.00000000	0	5	11
:	84.00000000	0	1	4
:	85.00000000	0	2	5
:	86.00000000	0	6	5
:	87.00000000	0	3	5
:	88.00000000	0	4	4
:	89.00000000	0	2	5
:	90.00000000	0	1	4
:	91.00000000	0	4	4
:	92.00000000	0	4	5
:	93.00000000	0	2	6
:	94.00000000	0	7	8
:	95.00000000	0	1	4
:	96.00000000	0	1	4
:	97.00000000	0	1	4
:	98.00000000	0	1	4
:	99.00000000	0	1	4
:	100.00000000	0	1	4
:	101.00000000	0	1	4
:	102.00000000	0	2	4
:	103.00000000	0	4	6
:	104.00000000	0	7	10

FCPHA, FORD CALHOUN, INTRO PHYSIOLOGY, BIOL 3111 760921140105000

4.2. allosteric				
:	105.00000000	0	2	2
4.3. enzyme nomenclature				
:	106.00000000	0	3	5
:	107.00000000	0	1	6
:	108.00000000	0	1	5
:	109.00000000	0	1	5
4.4. classification				
:	110.00000000	0	1	5
:	111.00000000	0	2	5
:	112.00000000	0	4	6
5. cofactors				
:	113.00000000	0	4	4
:	114.00000000	0	3	5
:	115.00000000	0	4	5
:	116.00000000	0	6	5
:	117.00000000	0	4	7
:	118.00000000	0	2	6
:	119.00000000	0	6	6

Figure 6. Contents listing for item file FCPHA. This outline parallels the lecture in Introductory Physiology.

TWO SAMPLE ITEMS FOR A DEMONST

TWO SAMPLE ITEMS FOR A DEMONSTRATION OF THE COMPUTER PROGRAMS 7610042349150000

Item from a genetics item file of 550 items.

139.55700000	1273	3	5		:	The following, listed in ord colored; c(ch), chinchilla; c, albino. A colored rabbit chinchilla. The parental genotypes could
1	424	20	79	45	+	c+ c(ch) X c+ c(ch)
2	400	21	81	46	+	c+ c(ch) X c+ c(h)
3	238	80	60	45	-	c+ c(ch) X c(ch) c(ch)
4	270	81	62	43	-	c+ c(ch) X c(ch) c(h)
5	255	79	59	44	-	c+ c(h) X c(ch) c(ch)
6	449	20	82	43	+	c+ c(ch) X c+ c
7	250	82	59	43	-	c+ c X c(ch) c(ch)
8	260	78	43	64	-	none of the other answers ar

Item from a physiology item file of 242 items.

159.50000000	1500	5	7		:	An amino acid residue that c chains (R groups)
1	300	40	62	50	+	serine
2	325	40	63	49	+	tyrosine
					.	comments can be placed before
3	275	40	61	51	+	threonine
4	290	20	75	50	+	aspartic acid
5	310	20	72	50	+	glutamic acid
6	214	90	51	49	-	proline
7	220	70	72	30	-	tryptophan
8	200	90	52	49	-	isoleucine
9	235	90	51	48	-	leucine
10	194	85	50	50	-	valine
11	212	82	50	50	-	alanine
12	225	81	48	51	-	glycine

Figure 7. Analytical listing of item file FIG4. (This is a truncated version of the actual 132 character listing, truncation allowed easier reproduction of this report.)

```
SELECTIONS = 2;VERSIONS = 2;
ITEM_FILE = FIG4; TEST_FILE = FIG10;
REQUEST = 2 100 -200;
LINESIZE = 80;
TITLE = SPECIFICATION EXAMPLE;
END;
```

Figure 8. Specifications for a set of 4 tests (2 versions of each of 2 selections).

```
761007200919, SPECIFICATIONS SOUGHT
SELECTIONS = 2;
VERSIONS = 2;
ITEM_FILE = FIG4;
TEST_FILE = FIG10;
REQUEST = 2 100 -200;
LINESIZE = 80;
TITLE = SPECIFICATION EXAMPLE;
END;

761007200919, SPECIFICATIONS OBTAINED
TITLE = SPECIFICATION EXAMPLE
TEST_FILE = FIG10
ANALYSIS_FILE =
KEY_PRINT_FILE =
KEY_PUNCH_FILE =
SELECTIONS =          2
VERSIONS =          2
FIRST_KEY =           1
LINESIZE =         80
PAGESIZE =         65
TOP_LINE =           1
SELECT#RIGHT =            1
SELECT#WRONG =            4
RANDOM_NUMBER =      26091

761007200919, POSSIBLE ITEMS
REQUEST    FILE       ITEMS
      1  FIG4          100.00000000     200.00000000

761007200919, AVAILABILITY OF REQUESTED ITEMS
REQUEST    #REQUESTED    #AVAILABLE
      1             2             2

761007200930, SPECIFICATIONS SOUGHT
GROUPS PROCESSED =      1
```

Figure 9. ITSGEN operated on specifications of figure 8 to produce this specifications listing and tests of figure 10.

761007200919, SELECTION 1, VERSION 1, KEY 1, #ITEMS 2
SPECIFICATION EXAMPLE

1. The following, listed in order of dominance, are 4 alleles in rabbits: c+, colored; c(ch), chinchilla; c(h) himalayan; c, albino. A colored rabbit crossed to a colored rabbit produced 3 colored: 1 chinchilla. The parental genotypes could have been (1) none of the other answers are correct (2) c+ c(h) X c(ch) c(ch) (3) c+ c(ch) X c+ c(h) (4) c+ c(ch) X c(ch) c(ch) (5) c+ c X c(ch) c(ch)

2. An amino acid residue that contributes to H bonding between amino acid side chains (R groups) (1) proline (2) alanine (3) threonine (4) valine (5) isoleucine

761007200919, SELECTION 1, VERSION 2, KEY 2, #ITEMS 2
SPECIFICATION EXAMPLE

1. The following, listed in order of dominance, are 4 alleles in rabbits: c+, colored; c(ch), chinchilla; c(h) himalayan; c, albino. A colored rabbit crossed to a colored rabbit produced 3 colored: 1 chinchilla. The parental genotypes could have been (1) c+ c X c(ch) c(ch) (2) c+ c(ch) X c+ c(h) (3) c+ c(ch) X c(ch) c(ch) (4) c+ c(h) X c(ch) c(ch) (5) none of the other answers are correct

2. An amino acid residue that contributes to H bonding between amino acid side chains (R groups) (1) threonine (2) isoleucine (3) valine (4) alanine (5) proline

761007200919, SELECTION 2, VERSION 1, KEY 3, #ITEMS 2
SPECIFICATION EXAMPLE

1. The following, listed in order of dominance, are 4 alleles in rabbits: c+, colored; c(ch), chinchilla; c(h) himalayan; c, albino. A colored rabbit crossed to a colored rabbit produced 3 colored: 1 chinchilla. The parental genotypes could have been (1) c+ c(ch) X c+ c (2) c+ c(ch) X c(ch) c(ch) (3) c+ c(ch) X c(ch) c(h) (4) none of the other answers are correct (5) c+ c X c(ch) c(ch)

2. An amino acid residue that contributes to H bonding between amino acid side chains (R groups) (1) proline (2) glycine (3) aspartic acid (4) alanine (5) valine

761007200919, SELECTION 2, VERSION 2, KEY 4, #ITEMS 2
SPECIFICATION EXAMPLE

1. The following, listed in order of dominance, are 4 alleles in rabbits: c+, colored; c(ch), chinchilla; c(h) himalayan; c, albino. A colored rabbit crossed to a colored rabbit produced 3 colored: 1 chinchilla. The parental genotypes could have been (1) none of the other answers are correct (2) c+ c(ch) X c(ch) c(ch) (3) c+ c X c(ch) c(ch) (4) c+ c(ch) X c(ch) c(h) (5) c+ c(ch) X c+ c

2. An amino acid residue that contributes to H bonding between amino acid side chains (R groups) (1) glycine (2) valine (3) aspartic acid (4) alanine (5) proline

---

Figure 10. ITSGEN operated on specifications of figure 8 to produce these 4 tests (2 versions of 2 selections) and the specifications listing of figure 9.

```
KEY_PRINT_FILE = FIG14;
ITEM_FILE = BDGNA;
REQUEST = 6 5 12 17 23 31 41;
ITEM_FILE = BDGNB;
REQUEST = 3 97 111 112;
ITEM_FILE = BDGNC;
REQUEST = 6 118 135 148 165 175 184;
ITEM_FILE = BDGNF;
REQUEST = 3 417 436 449;
TEST_FILE = FIG13;
TITLE = FIRST SAMPLE TEST, GENETICS, FALL 1976;
END;
```

Figure 11. Specifications for 1 test.

```
760928232541, SPECIFICATIONS SOUGHT
KEY_PRINT_FILE = FIG14;
ITEM_FILE = BDGNA;
REQUEST = 6 5 12 17 23 31 41;
ITEM_FILE = BDGNB;
REQUEST = 3 97 111 112;
ITEM_FILE = BDGNC;
REQUEST = 6 118 135 148 165 175 184;
ITEM_FILE = BDGNF;
REQUEST = 3 417 436 449;
TEST_FILE = FIG13;
TITLE = FIRST SAMPLE TEST, GENETICS, FALL 1976;
END;

760928232541, SPECIFICATIONS OBTAINED
TITLE = FIRST SAMPLE TEST, GENETICS, FALL 1976
TEST_FILE = SAMPLE
ANALYSIS_FILE =
KEY_PRINT_FILE = KEY
KEY_PUNCH_FILE =
SELECTIONS =          1
VERSIONS =
FIRST_KEY =           1
LINESIZE =          132
PAGESIZE =           65
TOP_LINE =            1
SELECT#RIGHT =            1
SELECT#WRONG =            4
RANDOM_NUMBER =        9281
```

```
760928232541, POSSIBLE ITEMS
REQUEST   FILE      ITEMS
     1  BDGNA          5.00000000
     1  BDGNA         12.00000000
     1  BDGNA         17.00000000
     1  BDGNA         23.00000000
     1  BDGNA         31.00000000
     1  BDGNA         41.00000000
     2  BDGNB         97.00000000
     2  BDGNB        111.00000000
     2  BDGNB        112.00000000
     3  BDGNC        118.00000000
     3  BDGNC        135.00000000
     3  BDGNC        148.00000000
     3  BDGNC        165.00000000
     3  BDGNC        175.00000000
     3  BDGNC        184.00000000
     4  BDGNF        417.00000000
     4  BDGNF        436.00000000
     4  BDGNF        449.00000000

760928232541, AVAILABILITY OF REQUESTED ITEMS
REQUEST   #REQUESTED   #AVAILABLE
     1            6            6
     2            3            3
     3            6            6
     4            3            3

760928232604, SPECIFICATIONS SOUGHT
GROUPS PROCESSED =     1
```

Figure 12. ITSGEN operated on the specifications of figure 11 to produce this specifications listing and the test of figure 13.

760928232541, SELECTION 1, VERSION 1, KEY 1, #ITEMS 18
FIRST SAMPLE TEST, GENETICS, FALL 1976

1. Mendel's Law of Independent Assortment states that each parent passes on
(2) 9:3:3:1 ratio (3) random chromosomes (4) heterozygosity (5) random combin

2. In peas, smooth seed (S) is dominant over wrinkled (s) and yellow seed (
yellow seeds is testcrossed. The 100 progeny plants are all smooth yellow. What
(1) ssYy (2) SsYY (3) SSYy (4) insufficient information (5) SSYY

3. The environment influences (1) no human traits (2) most characteristic
(4) only non-inherited traits (5) only a few traits

4. If A- is epistatic to B and b, what phenotypic ratio will result from th
(3) 1:1:1:1 (4) 9:6:1 (5) 13:3

5. A man and wife are both heterozygous for a recessive autosomal gene whic
probability that the first child will be a girl with phenylketonuria? (1) 1/4

6. If a chi-square test yields p of 0.001 the hypothesis is (1) correct wi
(2) questionable and should be retested (3) incorrect (4) incorrect with a pro

7. All cells in each individual of a species have the identical number of
(4) ribosomes (5) chromosomes

8. Growth in higher organisms occurs primarily by (1) mitotic cell divisio
(5) cytoplasmic expansion

9. A diploid cell has two metacentrics and two acrocentrics. After one mito
metacentrics and how many acrocentrics? (1) 2 and 0 or 0 and 2 (2) 2 and 1 or

10. Stage of meiosis in which homologs are paired (1) anaphase II (2) telo
(5) anaphase II

11. A human cell in G1 has about 6?-6 micrograms of DNA. How many micrograms
(3) 3E-6 (4) 6E-6 (5) 24E-6

12. Microsporogenesis differs from megasporogenesis (1) 2 meiotic divisions
(3) takes place in flowers (4) more than one nucleus in cell at fertilization

13. Meiosis I nondisjunction of sex chromosomes in a human male would lead t
(2) no sex chromosomes (3) 2 X's (4) 1 X (5) 1 Y

14. A man has ichthyosis, an X-linked recessive. His wife is homozygous norm
(1) 1/4 of children (2) none (3) 1/2 of children (4) all daughters (5) 1/2 s

15. Claret-eyed female flies are crossed to red-eyed males. All progeny are
(1) X-linked recessive (2) autosomal dominant (3) X-linked dominant (4) autos

---

Figure 13. ITSGEN operated on specifications of figure 11 to produce this test. (This is a truncated version of the actual 132 character, 18 item test; truncation allowed easier reproduction of this report.)

760928232541, SELECTION 1, VERSION 1, KEY 1, #ITEMS 18
(1) 5 (2) 5 (3) 2 (4) 1 (5) 4 (6) 1 (7) 5 (8) 1 (9) 3
(10) 3 (11) 2 (12) 2 (13) 2 (14) 2 (15) 2 (16) 5 (17) 2
(18) 1

Figure 14. Key for test of figure 13.

TESTING STATION PROCEDURE

Report to the testing station with a #2 pencil, picture ID (student ID, driver's licence, etc.), and other materials required by your instructor. You will not be admitted without these materials. Under no circumstances will scratch paper, books, notes, pets, or other non-required material be allowed in the testing room.

INITIATION OF TEST

1. You must show the monitor a picture ID. He will check it against the class roll and loan you a test.

2. Write your name at the top of the answer sheet.

3. Black out your student ID on the answer sheet (item 1).

4. Do all scratch work on the test.

5. Consult the monitor for technical advice. The monitor will not answer any questions concerning the subject matter or wording of your test.

SPECIAL CODES AND INSTRUCTIONS FOR SOME CLASSES

6. Exponential numbers are in E notation; e.g., 1.3 times ten to the minus five is 1.3E-5.

7. When answering items requesting 'second digit only', use only the second digit of your computation; i.e., starting at the left and ignoring leading zeroes, take the second digit; e.g.,
if your computation yields 903.75, your answer is 0,
if your computation yields 1.987E-4, your answer is 9,
if your computation yields 1.0, your answer is 0,
if your computation yields .00567, your answer is 6.

ON COMPLETION OF THE TEST

8. Have the monitor exchange your answer sheet for your test key.

9. Return to your seat and study your test. This is your only chance to study your test as it will be destroyed.

10. Please note any technical problems on the back of your test. Misspelled words, difficult phrases or terms, and multiple correct answers are examples. Also specifically note other short-comings of the test; e.g., no questions on subject X, too many questions on subject Y. Please make the monitor aware of these notes when you turn in the test so that the monitor will route your test to your instructor. Any notes you make may improve subsequent tests.

11. On completion of your study and before leaving the room, you must return the key and the test to the monitor. Removing tests from the testing room or copying test items is a violation of the Honor Code.

Figure 15. Instructions on the use of the testing station given to all students the week before the first testing period.

Please give us some facts and opinions to help in evaluation of this course and the out-of-class testing system used for your multiple choice test.

OPTIONAL Enter your ID number as usual in the upper left corner of the answer sheet. Your instructor will never see your responses. This request is made to permit correlating responses with course achievement. (Do not write name.)

1. Were the multiple choice test questions appropriate for course content?

17	_	_	(1) very appropriate
52	_	_	(2) adequate
23	_	_	(3) somewhat deficient
8	_	_	(4) really off-base

2. How well were the multiple choice questions written; to what extent were they understandable?

0	_	_	(1) excellent
27	_	_	(2) good
45	_	_	(3) fair
28	_	_	(4) poor

3. For me this course is

35	19	7	(1) very difficult
50	60	61	(2) difficult
15	19	27	(3) about average
0	0	4	(4) easy
0	2	0	(5) very easy

4. The lowest grade I would be satisfied with in this course is

15	15	14	(1) A
22	62	45	(2) B
33	23	38	(3) C
30	0	3	(4) D

5. For me this course is

14	13	5	(1) boring
16	16	8	(2) slightly boring
28	38	39	(3) slightly interesting
42	33	48	(4) interesting

6. Rate the overall quality of instruction in this course as compared with others you have taken at Virginia Tech.

7	5	2	(1) Poor
38	45	13	(2) Fair
33	42	58	(3) Good
22	8	28	(4) Excellent

The following questions concern only the out-of-class method of testing used for your multiple choice test, not the test questions or course.

7. The multiple choice test could be taken over a four-day period. Which day did you take it?

5	11	22	(1) first
13	15	12	(2) second
22	8	36	(3) third
60	64	29	(4) fourth
_	2	1	(5) cannot remember

For 8. thru 12: Various things may have influenced your choice of day. Tell whether each was an influence.

8. Tests scheduled in other courses

87	87	79	(1) yes
13	13	21	(2) no

9. Desire to gain as much study time as possible

80	74	59	(1) yes
20	26	41	(2) no

10. Illness

8	8	3	(1) yes
92	92	97	(2) no

11. Personal or social reasons

42	40	35	(1) yes
58	60	65	(2) no

12. Gain information from someone who took the test earlier

10	_	_	(1) yes
90	_	_	(2) no

13. Do you think you were able to do better on the multiple choice test (on the first try) as a result of choosing your own day?

34	68	67	(1) yes
29	16	11	(2) no
37	16	22	(3) cannot say

14. Did you retake the multiple choice test?

49	47	57	(1) yes
51	53	43	(2) no

15. Did availability of the retest influence how hard you studied for the first multiple choice test?

32	37	37	(1) yes
68	55	53	(2) no
_	8	10	(3) cannot say

16. If you retook the multiple choice test, do you feel you learned more as a result?

63	78	88	(1) yes
23	12	7	(2) no
14	10	5	(3) cannot say

17. Did the atmosphere of the testing room and freedom from time pressure aid your performance on the multiple choice test or retest compared with that of regular classroom testing?

54	81	81	(1) yes
23	13	15	(2) no
23	6	3	(3) cannot say

18. Which multiple choice testing approach do you prefer?

29	24	7	(1) inclass (all students take the same test on the same day)
71	76	93	(2) out-of-class (as in this class with opportunity to retake)

19. Suppose that the opportunity to retake was not available. Then which approach would you prefer?

50	42	20	(1) inclass
50	58	80	(2) out-of-class

20. Rate the helpfulness of receiving the key immediately after completing the multiple choice test.

43	39	57	(1) very helpful
39	46	37	(2) helpful
18	15	5	(3) not helpful

21. Do you feel that there was a substantial variation in degree of difficulty from one person's examination to another's?

_	22	16	(1) yes
_	22	24	(2) no
_	56	59	(3) cannot say

22. Did spreading examination time over a two-week period interfere with learning new material?

_	24	15	(1) yes, somewhat
_	16	36	(2) yes, slightly
_	60	50	(3) no

23. Write any comments or suggestions not covered above on the back of the opscan sheet.

Figure 16. Survey of three classes. The three columns are the percentage (rounded to integers) of students in each of the three classes that considered the item and marked the response.

"An excellent system."

"This system of testing seems exceptionally beneficial for courses such as genetics and physiology, because the student was given time to assimilate all the material."

"Overall, I like the system, and I definately perform and learn more through this method."

"Testing procedure in this individualized manner was very effective as a learning device. It was more relaxed, a more efficient use of time and more considerate of the student."

"I think this is a great system and wish other teachers would do the same."

"I think the retake system of exams is a excellent system. Any faults it has are a result of the short time period the quarter system gives. I sincerely feel that I learned much more from this course as a result and furthermore I wish all my classes had it."

"I prefer this system of testing because it makes you aware of areas you are weak in and gives you a chance to correct your weaknesses. I found it gave me more thorough knowledge of the material, and that's what I'm here for."

"I like it. Let's do more this way."

"Reduced pressure from not having a time factor involved really helped. Knowledge that a retest was available also aided in reducing pressures. Individual testing reduced amount of distractions. I believe all of the responses were helpful in not only doing better, but it is also a better indications of what is really learned."

"This new system really takes the pressure out of taking tests. I was able to study more effectively and did much better than usual. I hope this system catches on in the other departments."

"By being able to retake the exam we are able to know what type of questions will be asked and can be better prepared. Choosing our own day to take the exam avoids conflicts with other tests."

"I like the non pressurized testing procedure and the retake is of little concern, because I feel that I will do as well the first time as I would the second time when it comes to a 'computer' test like this. The choice of day is also very good and allows the student to schedule the test around pressing engagements."

Figure 17. Relevant positive student comments on item 23 of figure 16.

"Great system but (instructor) should be there to answer specific (course) questions."

"Teacher should be present to answer questions on ambiguity of some questions on the test."

Figure 18. Relevant negative student comments on item 23 of figure 16.

Table 1. Specifications for controlling generation of tests.

keyword	range (defaults parenthesized), description
control of test format	
LINESIZE	64-(132). Width of test page in characters.
PAGESIZE	40-(65). Length of test page in lines.
TOP_LINE	(1)-30. First print line of test page.
control of test identification	
FIRST_KEY	(1)-Y. Each test, and all material related to it, is uniquely and serially identified with a 'KEY' integer beginning with the value of FIRST_KEY.
TITLE	(blank) or 1-32,767 characters. Appears at the top of first page of each test, usually used for special instructions or identification. A title cannot contain a semicolon.
control of test items	
ITEM_FILE	1-7 character item file name must be specified.
RANDOM_NUMBER	(computed) or 5 digit odd number. Initial random number for pseudo random number generator. Two test generations will be equivalent if their RANDOM_NUMBER specifications are the same (assuming the same requests are made from the same item files).
REQUEST	must be specified. Number of items to be randomly selected followed by the IDs of the possible items; e.g., 'REQUEST = 10 17.5-39 73 97.6 107.7-3000.1;' requests a random selection of 10 items from a pool of items containing items 17.5 through 39, item 73, item 97.6, and items 107.7 through 3000.1. A REQUEST specification is associated with the first preceding ITEM_FILE specification. Any number of REQUEST specifications can be associated with an ITEM_FILE specification. In general, education is better served if each request is for 1 item from a homogeneous pool.

keyword	range (defaults parenthesized), description
control of test items continued	
SELECTIONS	(1)-Y. Number of tests to be generated by randomly selecting items and randomly selecting responses from the selected items.
SELECT#RIGHT	(1)-32,767. Maximum number of true responses to be randomly selected from each selected item.
SELECT#WRONG	1-(4)-32,767. Maximum number of false responses to be randomly selected from each selected item.
TEST_FILE	1-7 character file name must be specified. ITSGEN writes the tests on this print file.
VERSIONS	(1)-Y. Number of tests to be generated by randomizing the order of each item's responses. Thus, 2 versions of a selection are informationally identical.
control of auxiliary information	
ANALYSIS_FILE	(blank) or 1-7 character file name. ITSGEN writes all item and response IDs for each test generated on this file for subsequent use by ITSUPD.
KEY_PRINT_FILE	(blank) or 1-7 character file name. ITSGEN writes the test keys on this print file.
KEY_PUNCH_FILE	(blank) or 1-7 character file name. ITSGEN writes the test keys on this punch file.

Y must be consistent with 32,768 = SELECTIONS X VERSIONS + FIRST_KEY.

Table 2. The percentage (rounded to integers) of the 'right' class students who obtained the indicated pair of grades at their first testing period (N = 93).

test	retest				
	F	D	C	B	A
F	14	10	8	4	2
D	1	4	7	7	4
C	0	2	5	15	9
B	0	0	0	2	7
A	0	0	0	0	0